Name:.....

Data Mining and Visualization

Assignment 4

Classification

Assignments constitute part of the examination and must be handed in time. You are asked to hand in the solutions during a week following the week on which the assignment has been discussed in classes. You should submit your work to CANVAS.

Problem 1 – The optimal solution to the classification problem In this part we discuss the optimal solution to the classification problem with misclassification costs. Review corresponding material from the lecture. The goal is to provide argument that the following classification rule R minimizes the cost

$$R(\mathbf{x}) = \begin{cases} 0; & \dfrac{P(Y=0|\mathbf{x})}{P(Y=1|\mathbf{x})} > \dfrac{c(0|1)}{c(1|0)} \\ 1; & \dfrac{P(Y=1|\mathbf{x})}{P(Y=0|\mathbf{x})} > \dfrac{c(1|0)}{c(0|1)} \end{cases}$$

In this sense, this is the optimal rule for given classification costs $c(0|1)$ and $c(1|0)$.

1. Given that the observed $\mathbf{X}$ is a random variable any decision rule R is also a random variable taking value either 0 or 1. The observed variable Y is also such a variable but that, in principle, will not be observable (except for the training and testing data in supervised learning approach).
   (a) Represent in a form of two by tw0 table the joint distribution of (R, Y).
   (b) Express the probability of making a classification error while using the rule defined by R using the entries of the table of the joint distribution of (R, Y).
   (c) Identify two types of error and relate them to *specificity* and *sensitivity* of R.
2. Write the probability of a classification error using the conditional probability of R given Y.
3. Consider that the rule is completely non-random (i.e. no matter what you observe you decide either always for one or always for zero). Which of the possible two choices will give you the smallest error?
4. Consider an observable variable $\mathbf{X}$ and decision rules that are entirely decided based on the observed value $\mathbf{x}$ of $\mathbf{X}$, i.e. for some set G of $\mathbf{x}$'s:

   $$R = \begin{cases} 1 : \mathbf{x} \in G \\ 0 : \mathbf{x} \notin G \end{cases}$$

   Write the probability of error of misclassification given that $\mathbf{X} = \mathbf{x}$ is observed using the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$.
5. Use the argument from part 3 but with conditional probability given $\mathbf{x}$ to show that the optimal rule (in the terms of the error of misclassification) is given by

   $$R(\mathbf{x}) = \begin{cases} 0; & \dfrac{P(Y=0|\mathbf{x})}{P(Y=1|\mathbf{x})} \geq 1 \\ 1; & \dfrac{P(Y=0|\mathbf{x})}{P(Y=1|\mathbf{x})} < 1 \end{cases}$$

6. Let C be a random variable that describes the cost of a decision rule R. Describe the distribution of C as a function of the joint distribution of R and Y.
7. Express the expected value of C (the expected value of the classification rule R) in the terms of the joint probabilities of R and Y.
8. Introduce the conditioning on $\mathbf{x}$. Can you express the expected value of C using the conditional distribution on $\mathbf{X} = \mathbf{x}$?
9. Extend previous argument to prove the optimality of $R(X)$ with cost of classification?
10. Rewrite the results using the conditional distribution of $\mathbf{X}$ given Y.

Problem 2 – normal distribution, equal variances Suppose that X has normal distribution with variance $\sigma^2 = 1$ and the mean $\mu_1 = 1$ or $\mu_0 = 2$, depending if $Y = 1$ or $Y = 0$.

1. With no cost of misclassification and assuming that the prior probabilities for Y are equal derive the optimal classification rule based on the previous problem.
2. If you observe $X = 0.5$, how would you classify according to your rule.
3. Repeat previous parts with prior probability of $Y = 1$ being twice of that of $Y = 0$.

Problem 3 – normal distribution, non-equal variances Suppose that X has normal distribution with variances $\sigma_1^2 = 1$ and $\sigma_0^2 = 4$ and the mean $\mu_1 = 1$ or $\mu_0 = 2$, depending if $Y = 1$ or $Y = 0$.

1. With no cost of misclassification and assuming that prior probabilities for Y are equal derive the optimal classification rule as discussed in the first problem.
2. If you observe $X = 0.5$, how would you classify according to your rule.
3. Repeat previous parts with prior probability of $Y = 1$ being twice of that of $Y = 0$.

Problem 4 – classification based on the training sample Suppose that you have two samples, the first corresponding to $Y = 0$:

$$0.64, 1.75, 0.91, 0.46, 1.81, 1.13$$

and the second corresponding to $Y = 1$:

$$6.07, 2.37, 4.38, 0.95, 4.75, -2.50.$$

1. Explain how would you approach to the classification problem based on this sample assuming that the underlying distributions are normal and the prior probabilities are the same.
2. Derive the classification rule.
3. Classify according to the following values of x:

$$4.11, -1.89, 6.75, 3.15, 2.67, 0.67$$

4. Evaluate the probability of a misclassification error that is associated with your rule.